Statistical Analysis of Word Frequency Distribution in Texts of Different Genres: Comparison of Lithuanian and English

J. Mandravickaitė[1, 2], T. Krilavičius[2, 3]

[1] Vilnius University
[2] Baltic Institute of Advanced Technology
[3] Vytautas Magnus University
justina@bpti.lt

We report an ongoing study on statistical characteristics of texts written in different genres. It has been suggested that genres resonate with people because they provide familiarity and the shorthand of communication. Also, genres tend to shift hand-in-hand with public opinion and reflect widespread culture of certain period(s). From NLP perspective, genres come in use in text classification and categorization, natural language generation, etc.

At this stage, we present a statistical analysis of Lithuanian and English texts of genres. For our explorations, we use Corpus of the Contemporary Lithuanian Language (for Lithuanian part) and Freiburg-LOB Corpus of British English (F-LOB). The main points of interest are number of words, number of different words and word frequencies. Structural type distribution and Zipf's law were applied in order to describe the frequency distribution of words in different textual genres.

Zipf's law is one of the universal laws proposed to describe statistical regularities in language. Thus word frequencies and their derivative indicators could be used to characterize textual genres. Application of word rank-frequency distribution, type-token ratio, the percentage of hapax legomena, i.e., words that occur only once, and entropy for different genre groups (fiction, scientific articles, documents, news articles) supported the latter assumption.

Differences between languages (Lithuanian and English) were observed as well. As genres are rather complex phenomena that depend on various linguistic, cultural, societal, etc. factors, our future study includes research of additional frequency structure indicators as well as their combinations.